Resumo do capítulo - Organização de dados

Clube do Livro - R Para Ciência de Dados (2ª Edição)

Angélica Custódio

2024-08-06

Preparação: Pacotes

Esse capítulo foca bastante no *tidyr*, que faz parte do grupo de pacotes do *tidyverse*.

```
# Instalando "tidyverse"
install.packages("tidyverse")

# Instalando "dados"
remotes::install_github("cienciadedatos/dados")
```

“Dados bagunçados são bagunçados à sua maneira”

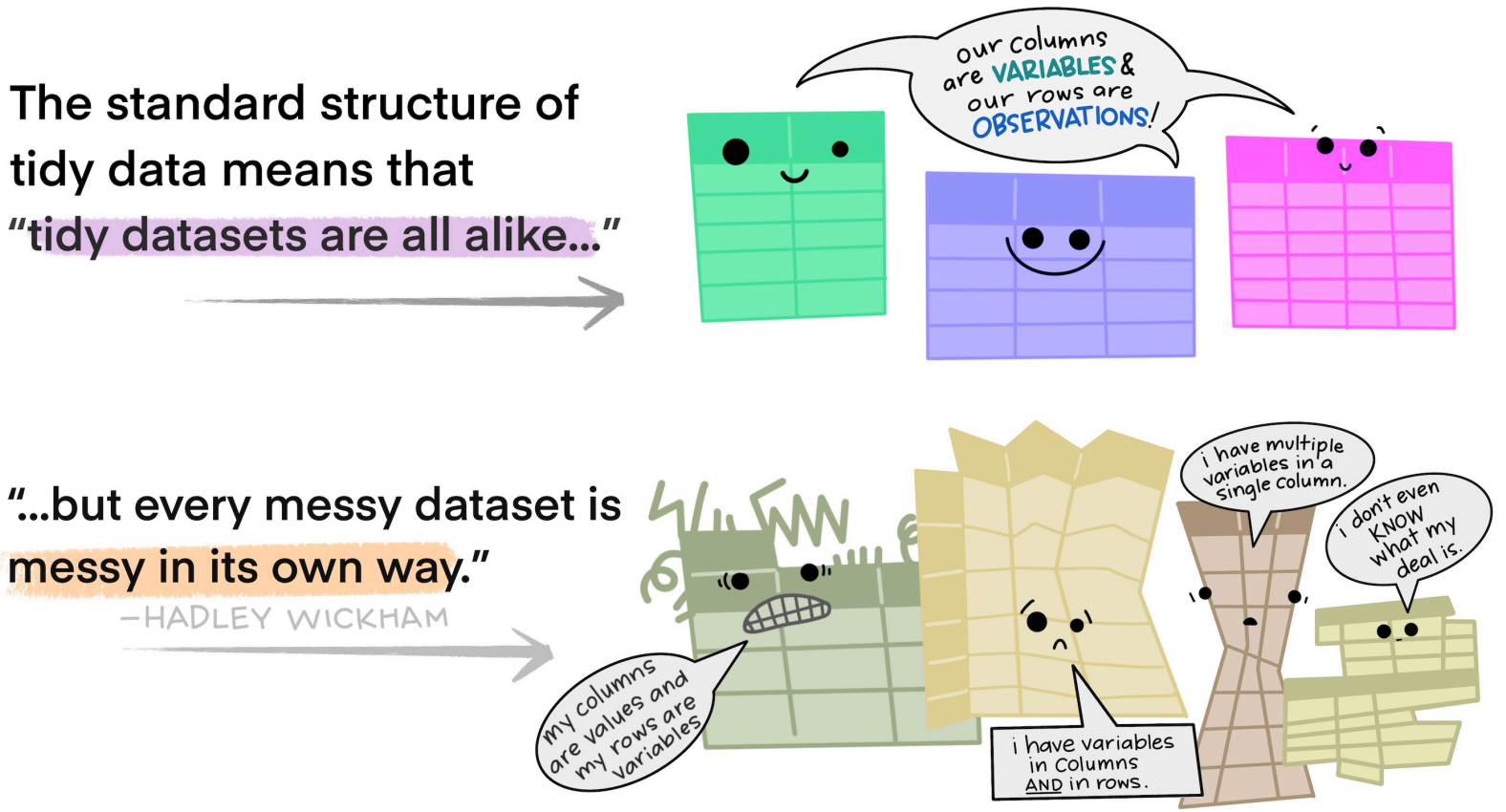


Ilustração - Allison Horst

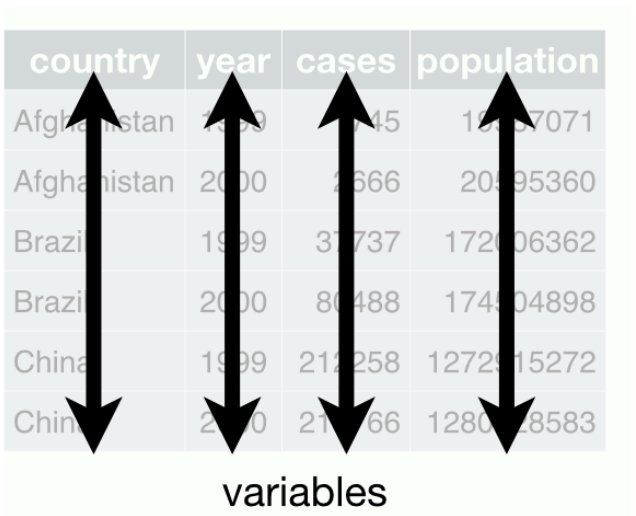
Dados *tidy*

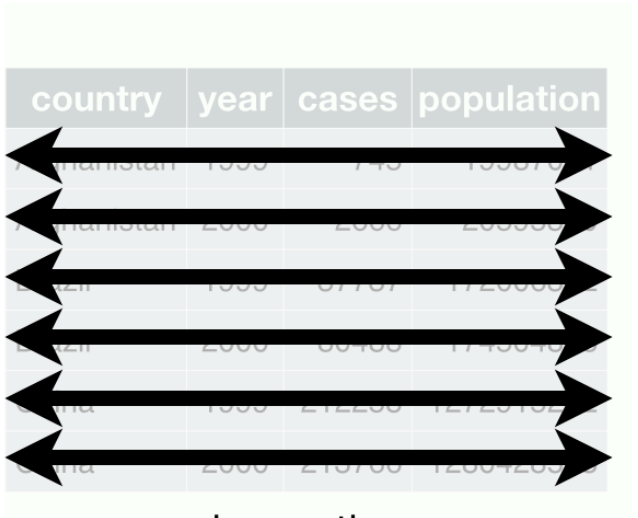
Existem três regras interrelacionadas que fazem com que um conjunto de dados seja considerado tidy:

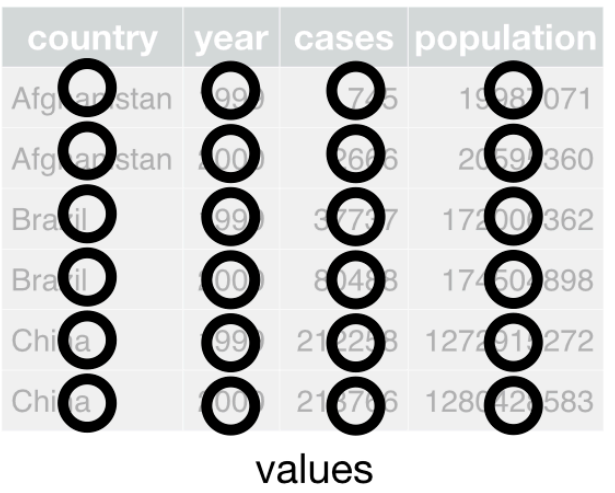
- Cada variável é uma coluna; cada coluna é uma variável.
- Cada observação é uma linha; cada linha é uma observação.
- Cada valor é uma célula; cada célula é um único valor.

dplyr, *ggplot2* e os demais pacotes do tidyverse foram pensados para trabalhar com *dados tidy*.

Dados *tidy*


country
year
cases
population
variables


country
year
cases
population
observations


country
year
cases
population
values

Exercícios

Abordando os exercícios nessa seção.

1. Para cada uma das tabelas do exemplo, descreva o que cada observação e cada coluna representa.

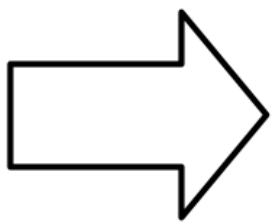
Exercícios

2. Faça um esboço do processo que você usaria para calcular taxa para a tabela2 e tabela3. Você precisará executar quatro operações:

- Extrair o número de casos de tuberculose por país por ano.
- Extrair a população correspondente por país por ano.
- Dividir os casos pela população e multiplicar por 10000.
- Armazenar de volta no local apropriado.

pivot_longer()

id	bp1	bp2
A	100	120
B	140	115
C	120	125



id	measurement	value
A	bp1	100
A	bp2	120
B	bp1	140
B	bp2	115
C	bp1	120
C	bp2	125

pivot_wider()

Utilizá-lo torna o conjunto de dados mais largo (wider), aumentando o número de colunas e diminuindo o número de linhas.

É muito útil quando uma informação está espalhada em múltiplas linhas.

Como definir a versão tidy?

Pode ser impossível especificar se a versão longa (long) ou larga (wide) é a versão tidy.

O recomendado é seguir com uma tabela organizada que faça sentido para o uso necessário.

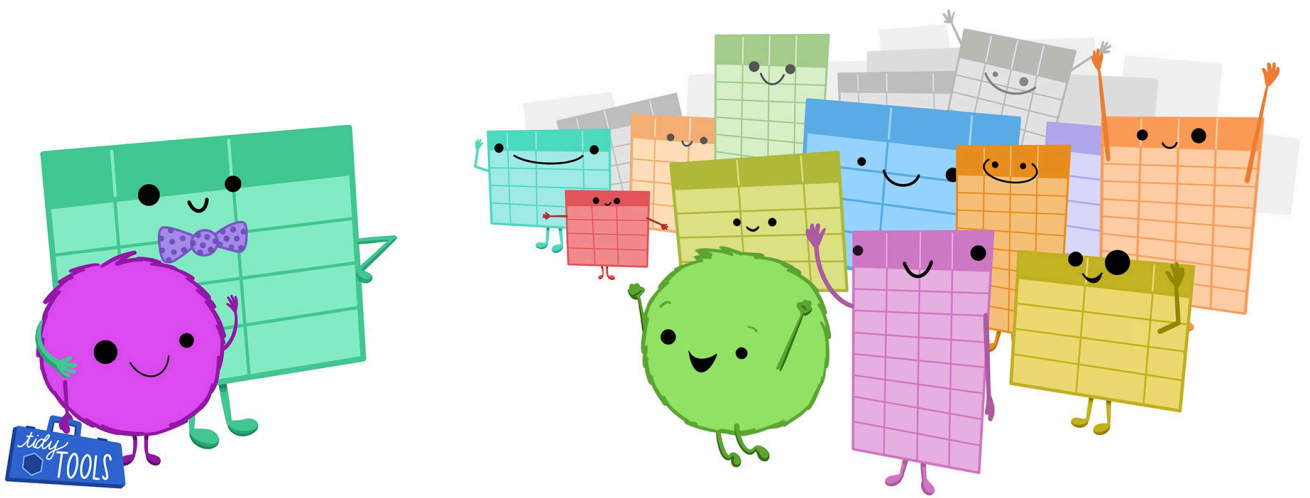


Ilustração - Allison Horst

Seguimos felizes com os dados organizados

Ilustração - Allison Horst

Referências

- Wickham, H. . (2014). Tidy Data. Journal of Statistical Software
- Data tidying with tidyr::Cheatsheet
- Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration - by Julia Lowndes and and Allison Horst